

Visual Inference by Data-Driven Markov Chain Monte Carlo

Zhuowen Tu and Song-Chun Zhu

Statistics and Computer Science
University of California, Los Angeles

Los Alamos National Lab, 12-2-2002

Parsing Image Into Various Stochastic Patterns

Depending on the types of patterns it focuses, image parsing subsumes conventional vision tasks:
perceptual organization, image segmentation, object recognition, etc.

Los Alamos National Lab, 12-2-2002

A Bayesian Formulation

A basic assumption, dated back to Helmholtz (1860), is that biologic and machine vision is to compute the most probable interpretation(s) from input images.

Let I be an image and W be a semantic representation of the world.

$$W^* = \arg \max_{w \in \Omega} p(W | I) = \arg \max_{w \in \Omega} p(I | W)p(W)$$

In statistics, we sample from a posterior probability to preserve ambiguities.

$(W_1, W_2, \dots, W_k) \sim p(W | I)$

Los Alamos National Lab, 12-2-2002

Problems

1. Representational or modeling problems:
What are W , $p(W)$, and $p(I | W)$?

2. Computational problems:
a). What are the structures of the search space, which we call Ω ?
b). How do we explore the search space for **globally optimal solutions**?
... reversible MC jumps + diffusion (PDEs).
c). How do we compute and preserve **ambiguities**.

Can MCMC run in seconds on a PC for parsing images?

Los Alamos National Lab, 12-2-2002

Ideas to Improve MCMC Speed in Literature

A main idea is to introduce auxiliary random variables:

$$x \sim \pi(x)$$

Augment x by variables:

- T --- temperature (Simulated tempering, Narinari and Parisi, 92, Geyer and Thompson, 95)
- s --- scale (Multi-grid sampling, Goodman and Sokal 88, Liu et al 94)
- w --- weight (dynamic weighting, Liang and Wong 1996)
- b --- bond (clustering, Swendsen-Wang, 87)
- u --- energy level (slice sampling, Edwards and Sokal, 88 ...)

The common problem is:
The Markov chain moves are designed a priori, without looking at the data.

Los Alamos National Lab, 12-2-2002

What is Data-Driven Markov Chain Monte Carlo ?

The complexity of sampling the posterior $W \sim p(W | I)$ is in the Metropolis-Hastings jumps

Consider a reversible jump $W_A \leftrightarrow W_B$

$$\alpha(W_A \rightarrow W_B) = \min(1, \frac{p(W_B | I)G(W_B \rightarrow W_A)}{p(W_A | I)G(W_A \rightarrow W_B)}) \text{ or } \min(1, \frac{p(W_B | I)q(W_A | W_B)}{p(W_A | I)q(W_B | W_A)})$$

Without looking at the data, the pre-designed proposal probabilities are often uniform distributions, thus it is a blind (exhaustive) search !

In DDMCMC,

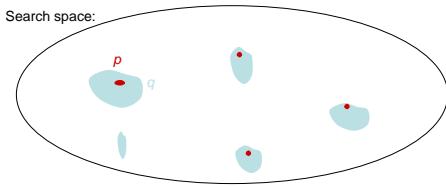
$$\alpha(W_A \rightarrow W_B) = \min(1, \frac{p(W_B | I)q(W_A | W_B, I)}{p(W_A | I)q(W_B | W_A, I)})$$

If $q(W_B | W_A, I) \equiv p(W_B | I)$, $q(W_A | W_B, I) \equiv p(W_A | I)$

Then it may converges in a small number of steps !

Los Alamos National Lab, 12-2-2002

Basic Ideas



The proposal probabilities $q()$ focuses on a tiny portion of the search space and thus narrows the search exponentially in a probabilistic fashion. Thus the Markov chain converges and mixes very fast.

Los Alamos National Lab, 12-2-2002

Intuitive Idea: Divide-and-Conquer

Let $W=(w_1, w_2, \dots, w_n)$, usually these variables are divided for several types:

partition, label of models, model parameters, order, ...

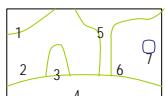
Consequently, the search space is made of a few types of "atomic spaces" --- one for each type of variables --- through union and production.

Then we can compute discriminative probabilities in each atomic space, which is then composed into the proposal probabilities.

Los Alamos National Lab, 12-2-2002

Example: Image Segmentation

$$W = (n, \{R_i, l_i, ?_i : i = 1, 2, \dots, n\}) \in \Omega$$



$p_7 = (R_1, R_2, \dots, R_7)$ is a 7-partition of the lattice.

$$V_{p_n} = \{ p_n = (R_1, R_2, \dots, R_n) : \bigcup_{i=1}^n R_i = \Lambda, R_i \cap R_j = \emptyset, \forall i \neq j \} / p_g$$

The partition space is $V_p = \bigcup_{n=1}^{|A|} V_{p_n}$ A permutation group

Los Alamos National Lab, 12-2-2002

Some Image Models

Some families of image models:

\mathbf{v}_{g1} : iid Gaussian for pixel intensities

\mathbf{v}_{g2} : non-parametric histograms

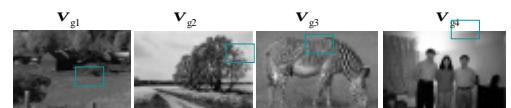
\mathbf{v}_{g3} : Markov random fields for texture

\mathbf{v}_{g4} : Spline model for lighting variations

\mathbf{v}_{c1} : iid Gaussian for color (LUV)

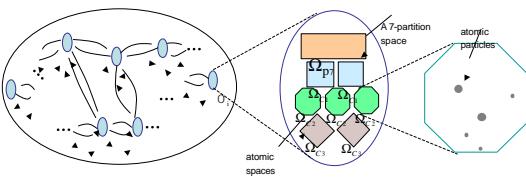
\mathbf{v}_{c2} : mixture of Gaussians for color

\mathbf{v}_{c3} : spline model for smooth color variations (e.g. sky, lake, ...)



Los Alamos National Lab, 12-2-2002

The Search Space



Los Alamos National Lab, 12-2-2002

Designing Markov Chain Dynamics

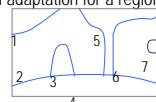
Type I: Diffusion of region boundary -- region competition.

Type II: Splitting of a region into two.

Type III: Merging two regions into one.

Type IV: Switching the family of models for a region.

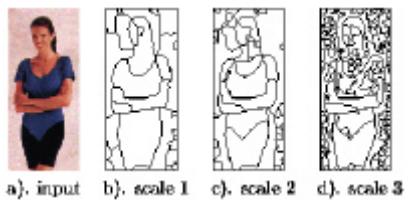
Type V: Model adaptation for a region.



Los Alamos National Lab, 12-2-2002

Edges in Partition Space \mathbf{V}_p

Edge detection and tracing at three scales of details:



a). input b). scale 1 c). scale 2 d). scale 3

Los Alamos National Lab, 12-2-2002

Clustering in Color Space \mathbf{V}_{c1}

Mean-shift clustering (Cheng, 1995, Meer et al 2001)

$$q(\cdot | I) = \sum_{i=1}^K \cdot_i g(\cdot - \cdot_i)$$

Input



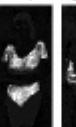
saliency maps



1



2



3



4



5



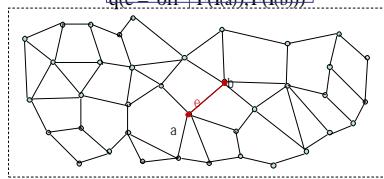
6

The brightness represents how likely a pixel belongs to a cluster.

Los Alamos National Lab, 12-2-2002

Walking in the Partition Space

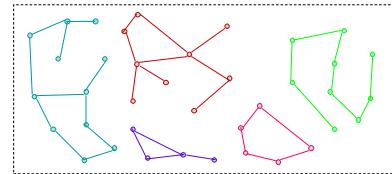
an adjacency graph: each vertex is a basic element : pixels, small-regions, edges, each link $e=a, b$ is associated with a probability/ratio for similarity
 $q(e = "on" | F(I(a)), F(I(b)))$
 $q(e = "off" | F(I(a)), F(I(b)))$



Los Alamos National Lab, 12-2-2002

Walking in the Partition Space

Sampling the edges independently, we get connected components:

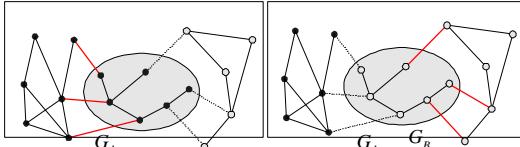


These connected sub-graphs are the **clusters** in the partition space

sampling $c \sim q(C | F(I))$ on \mathbf{V}_p

Los Alamos National Lab, 12-2-2002

Graph Partitioning– Generalizing SW



The red edges are the bridges $E(V^c, V_l - V^c), E(V^c, V_r - V^c)$

Theorem. Accepting the label change proposal with probability:

$$\alpha = \min \left\{ 1, \frac{p(B)}{p(A)} \frac{p(l|V^c, B, G) \prod_{e \in E(V^c, V_l - V^c)} (1 - q_e)}{p(A) p(l'|V^c, A, G) \prod_{e \in E(V^c, V_l - V^c)} (1 - q_e)} \right\}$$

results in an ergodic and reversible Markov Chain.

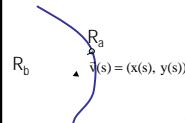
Los Alamos National Lab, 12-2-2002

Diffusion Components by PDEs

The Markov chains realized reversible jumps between sub-spaces of varying dimensions.

Within a subspace of fixed dimension, there are various diffusion processes expressed as partial differential equations.

For example, the **region competition** for curve evolution (Zhu, Lee, and Yuille, 95)



Let v be a point on the boundary between two regions, its motion is governed by the region-competition equation.

$$\frac{dv(s)}{dt} = (\mu \cdot ?_a(s) + \frac{\log p(I(x, y) | ?_a)}{\log p(I(x, y) | ?_b)}) \cdot \vec{n}(s)$$

Los Alamos National Lab, 12-2-2002

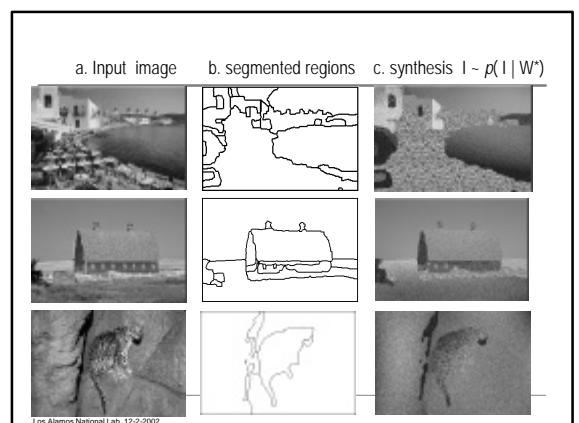
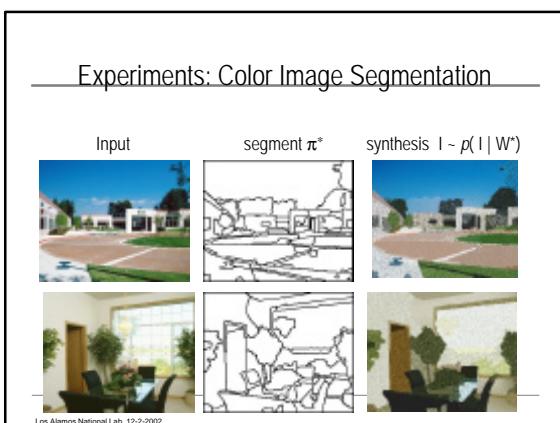
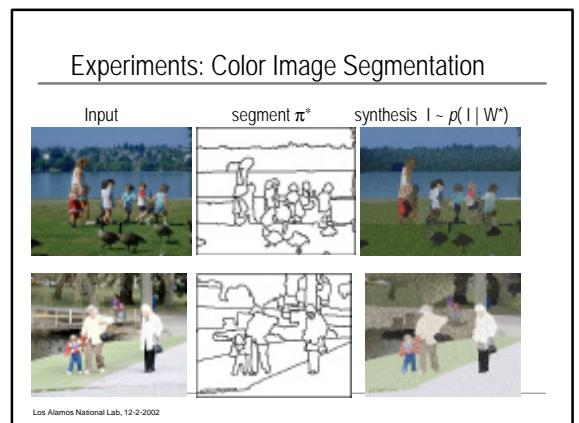
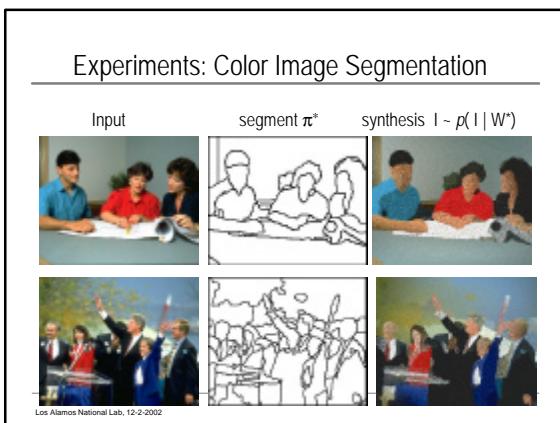
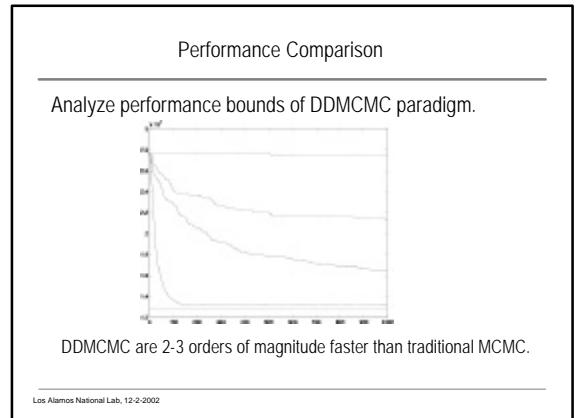
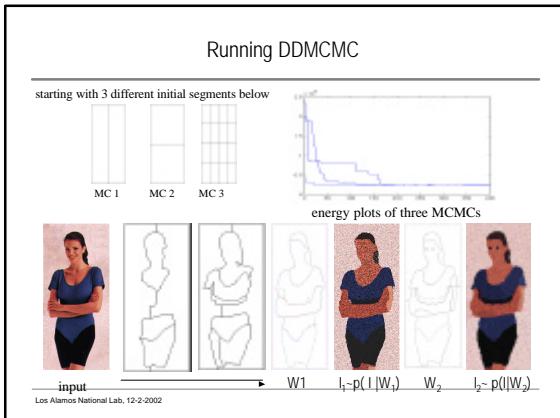


Image Segmentation

Input segment π^* synthesis $I \sim p(I | W^*)$

Los Alamos National Lab, 12-2-2002

The Berkeley Benchmark Study (David Martin et al, 2001)

test images	DDCMC	manual segment	"error" measure
			0.1083
			0.3082
			0.5627

Los Alamos National Lab, 12-2-2002

Examples of Failure

a. Input image b. segmented regions c. synthesis $I \sim p(I | W^*)$

Los Alamos National Lab, 12-2-2002

Adaboost in the Label Space

--- an example from Viola and Jones, 2001.

a. the first two face features

b. an example of face detection

Adaboost is a learning algorithm which makes decision by combining a number of simple features. As T and training samplers become large enough, it weakly converges to the log ratio of the posterior probability

$$y = \text{Sign}(a_1 h_1(I) + \dots + a_T h_T(I)) \rightarrow \text{sign}(p(y=1|I)/p(y=-1|I))$$

Los Alamos National Lab, 12-2-2002

Ambiguities in Visual Inference

Nicker cube Vase vs. faces bikini vs. martini

Los Alamos National Lab, 12-2-2002

Ambiguity in Visual Inference

a. Input image b. Segmented texture regions c. synthesis by texture models

d. curve processes + bkgd region e. synthesis by curve models

Los Alamos National Lab, 12-2-2002

Computing Multiple Solutions

To faithfully preserve the posterior probability $p(W|\mathbf{I})$,

We compute a set of weighted scene particles $\{W_1, W_2, \dots, W_M\}$,

$$\hat{p}(W|\mathbf{I}) = \sum_{i=1}^M \alpha_i G(W - W_i), \quad \sum_{i=1}^M \alpha_i = 1$$

A mathematical principle:

$$S^* = \{W_1, W_2, \dots, W_M\} = \arg \min_S D(p||\hat{p})$$

Los Alamos National Lab, 12-2-2002

Pursuit of Multiple Solutions

$$D(p||\hat{p}) \approx \log \frac{\hat{w}}{w} + \sum_{i=1}^N \frac{w_i}{\hat{w}} \left[(E(\mathbf{x}_{i(t(n))}) - E(\mathbf{x}_n)) + \frac{(\mathbf{x}_n - \mathbf{x}_{i(t(n))})^2}{2\sigma^2} \right] = \hat{D}(p||\hat{p})$$

The Kullback-Leibler divergence can be computed if we assume mixture of Gaussian distributions.

— a simple fact: the KL-divergence of two Gaussians is the signal-to-noise ratio.

Intuition: S includes global maximum, local modes, apart from each other.

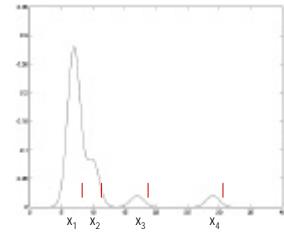
Los Alamos National Lab, 12-2-2002

A k -adventurer algorithm

1. Initializing S_k by one \mathbf{x} repeated k times and initializing $\hat{p}|\mathbf{x}$.
2. Repeat
3. Get k new weighted particle (w'_i, \mathbf{x}'_i) by k MCMCs starting at \mathbf{x}_i respectively.
4. $S_k \leftarrow S_k \cup \{(w'_i, \mathbf{x}'_i), i = 1, \dots, k\}$.
5. $p(\mathbf{x}) \leftarrow \hat{p}(\mathbf{x})$ by adding new particles.
6. $s^* = \arg \min_{s \in \{1, \dots, k+1\}} D(p||\hat{p}|\{s\})$.
7. $S_k \leftarrow s^*$.

Los Alamos National Lab, 12-2-2002

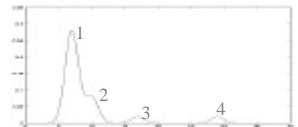
Preserving Distinct Particles



Los Alamos National Lab, 12-2-2002

An Example of Keeping Multiple Solutions

An example of illustration:



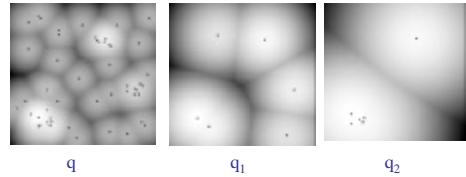
chosen S_0 :	$\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$	$\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_4\}$	$\{\mathbf{x}_1, \mathbf{x}_3, \mathbf{x}_4\}$	$\{\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$
$D(p \hat{p})$:	3.5487	1.1099	0.5373	2.9430
$\hat{D}(p \hat{p})$:	3.5487	1.1044	0.4263	2.8230
$ \hat{p} - p $:	0.1000	0.1000	0.3500	1.2682

Los Alamos National Lab, 12-2-2002

Preserving Distinct Particles

An example of illustration:

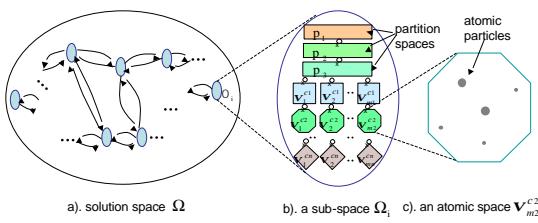
$$\min |q - q_1|$$



A model p with 50 particles two approximate models q with 6 particles

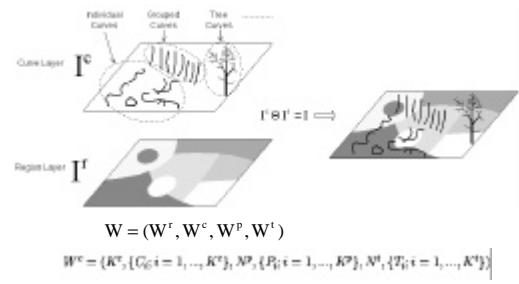
Los Alamos National Lab, 12-2-2002

General Search Space for Image Parsing



Los Alamos National Lab, 12-2-2002

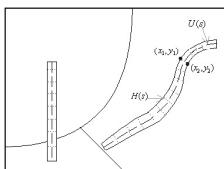
Parsing Images into Regions and Curves



Los Alamos National Lab, 12-2-2002

Curve Models

$$W^c = (K^c, \{(C_i, a_i); i=1..K^c\})$$



$$\text{Curve } C = (\Gamma, q)$$

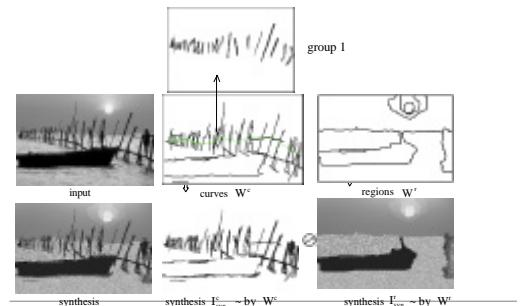
$$\text{Curve shape } \Gamma = (U(s), H(s))$$

$H(s)$ is the curve width.

$U(s)$ is the center line.

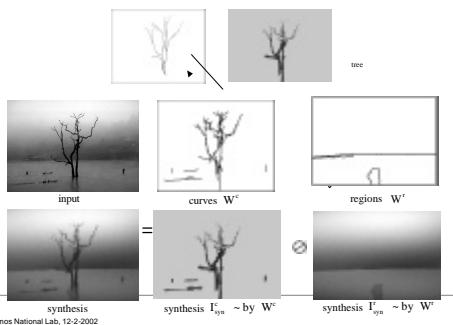
Los Alamos National Lab, 12-2-2002

Parse Image into Regions and Curves



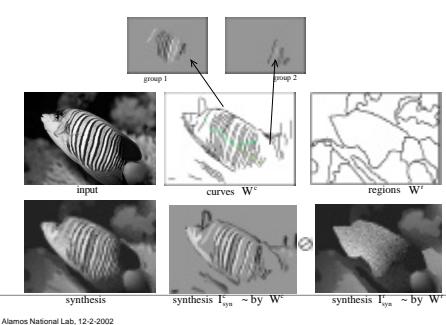
Los Alamos National Lab, 12-2-2002

Parsing Images with Trees



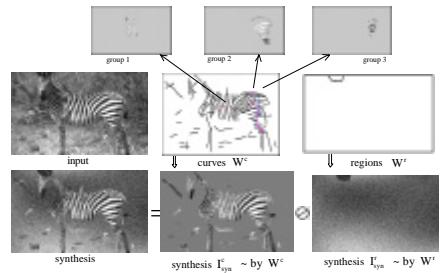
Los Alamos National Lab, 12-2-2002

Parse Image into Regions and Curves



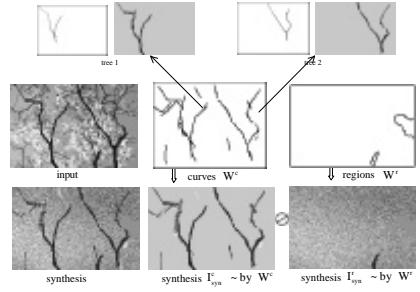
Los Alamos National Lab, 12-2-2002

Parse Image into Regions and Curves



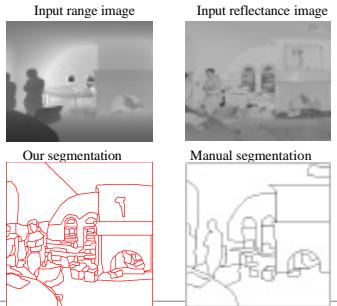
Los Alamos National Lab, 12-2-2002

Parsing Images with Trees



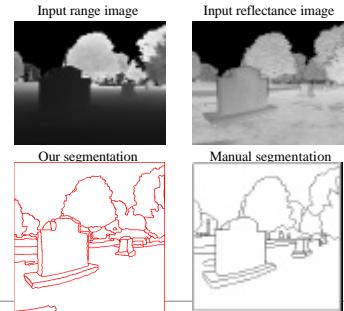
Los Alamos National Lab, 12-2-2002

Segmenting Laser Range Images



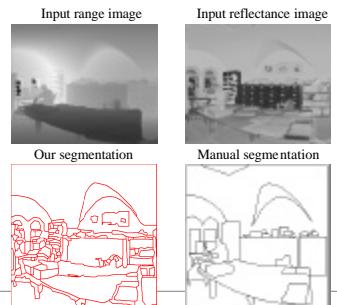
Los Alamos National Lab, 12-2-2002

Segmenting Laser Range Images



Los Alamos National Lab, 12-2-2002

Segmenting Laser Range Images



Los Alamos National Lab, 12-2-2002

Two Computing Paradigms in Vision

1. Generative methods ... "Top-down"

explicitly model the visual patterns

- Bayesian framework,
- Markov random fields,
- Markov chain Monte Carlo,
- Partial differential equations for diffusion, evolving, ...

◀ General but quite slow

2. Discriminative methods ... "Bottom-up"

explore "intra-class" vs "inter-class" difference

- Feature extraction, on/off, e.g. Edge detection
- Data clustering
- AdaBoost,
- Decision tree, ...

◀ Fast but not reliable

Summary

1. DDMCMC is a systematic way for integrating "top-down" and "bottom-up".

The discriminative methods approximate local posterior probabilities (ratios) in various atomic spaces. These probabilities/ratios are used as importance proposal probabilities, and drive the Markov chain to search for globally optimal solutions.

2. Fast Markov chain convergence and mixing at low temperature.

In contrast to simulated annealing, the SW-type algorithm can move fast at low temperature.

3. Ensemble complexity vs. worst case complexity

Though one can always construct worst case and prove NP-completeness, but on the average case, the computational complexity can be much lower.

Los Alamos National Lab, 12-2-2002

When the bottom-up proposal probabilities fail !



Los Alamos National Lab, 12-2-2002